EXPERT PANEL ON PUBLIC USE DATA ACCESS AND DISCLOSURE CONTROL

MEETING NOTES

(Version 2)

October 11, 2007
Westat, Rockville, MD

Purpose of the Panel

Because the NCS data are intended to become a national resource, the study hopes to make data accessible not only to NCS investigators, but to non-NCS investigators including adjunct study investigators, the general scientific community, and the general public. We must balance our obligations to science, to NCS investigators, and to the public, however, with an obligation to protect the confidentiality and privacy of the families who will generously share confidential information about their lives with the study. The unusual breadth of exposures and outcomes being collected at multiple time points makes this dataset highly vulnerable not only to identity disclosure but attribute and inferential disclosure as well. This expert panel was convened to share their experiences and expertise, and to offer recommendations on how to provide as much access as possible to investigators while insuring the confidentiality of NCS participants.

Attendees

Jennifer Madans, NCHS (Panel chair)
Myron Gutmann, University of Michigan
Marilyn Seastrom, NCES
Paul Sorlie, NHLBI
Alan Zaslavsky, Harvard University
Peter Scheidt, NICHD, NCS Program Office
Sarah Knox, NICHD, NCS Program Office
Alexa Fraser, Westat, NCS Coordinating Center
Marsha Hasson, Westat, NCS Coordinating Center
Nancy Weinfield, Westat, NCS Coordinating Center

1. Levels of Users

Levels of users, and proposals for access, release, and disclosure control were presented to the Panel. These levels appear in Table 1.

- The Panel recommended that even within a level of user, not everyone in the level should automatically have access to all possible data for that level.

1.1 NCS Community of Investigators – Levels 0 and 1

- The Panel was concerned about distribution of data within the NCS. For example, should young scientists have access to the data if they might have begun working with a PI on the NCS but subsequently move elsewhere?
- The definition of who is included in the NCS Investigators, or who qualifies as someone who is "internal" to the study, needs to be made clear both for policy and for the public so that people outside the study understand who is defined as an insider. This will be particularly important because data are being released at different times to the internal vs. external investigators.
- The NCS needs to consider whether we are suppressing the right data elements for data files WITHIN the NCS Community of Investigators – are we actually suppressing too much data when we de-identify?

- Data access and data ownership issues need to be clarified so that the relationship between the sponsor and the sites is clear and documented. The current contract stipulates that the government decides policies for privacy and confidentiality of the data, and that the SCs will agree to those policies. The government is ultimately the steward of the data, but details should be spelled out in policies.
- It will be important to clarify how data will be managed and overseen at the SCs and other internal sites of NCS Investigators.
- The Panel strongly encouraged the NCS to apply the same standards to the ICC that are applied to everyone else in the NCS. The ICC should remain under the same access control procedures. The data use agreements should be with individual investigators at the agencies, not with the agencies themselves.
- Data access security for the NCS collaborating agencies will have to be dealt with separately, and may be challenging. NCES does not require a security plan from other agencies, for example, because all the agencies they work with have the same certification and accreditation standards.
- Even within the NCS Community of Investigators the NCS should restrict access to "sensitive" data.
- Within an SC not everyone should require access to the full range of data. Access should be granted with permission and as needed.
- Although the Publications Subcommittee has said that some publications that are site or center specific are not acceptable (with the exception of approved adjunct studies), SCs may still wish to publish site-specific data once their data are available. The Panel suggested that such publications might focus on topics that do not endanger participant confidentiality in these smaller subsamples. Parameters of what can and cannot be done with these data should be made clear. The Stanford case was mentioned as legal precedent of concern with regard to setting vague limits on publication of data.
- If site- or center-specific analyses are permitted within the NCS Community of Investigators, the NCS should be very sensitive to confidentiality issues, as the risk of re-identification is much greater in such analyses than in the study-wide data.

1.2 Adjunct Studies – Level 2

- A more thorough plan is needed to ensure that adjunct studies' data get added to the main NCS data. The current system proposed by the Program Office is that the adjunct studies' investigators will turn the data from their additional data collection over to the CC. The CC will match their data with the rest of the NCS data and redistribute the adjunct study data with the approved NCS data to the sponsoring PI. Details of this plan, and the roles and responsibilities of all the involved parties, would need to be specified in advance.
- The Panel recommended that all adjunct study agreements should specify that adjunct study data will be retained by the NCS.
- Adjunct study policies should be very specific with regard to how long the adjunct study investigators get exclusive access to the adjunct data. Some studies give exclusivity for 6 months or less after the data are released to the adjunct study investigator.
- Adjunct studies may pose challenges for disclosure control because they will contain specific individual information and very circumscribed geographical scope.
- Within adjunct studies' investigators, as we have planned, access should be limited to the data needed for the proposed and approved analyses.
- The Panel wondered whether site-specific data should be issued at all. Most adjunct study data will be site specific, but those data will not contain the full complement of data collected – only data relevant to the adjunct study analyses. Panel members voiced concern about permitting adjunct studies that involve single sites due to this increased disclosure risk.

- According to the current plan, adjunct study investigators will have more access restrictions than the general scientific community, but the general scientific community will have more stringent disclosure control. To have this seem reasonable to outsiders the strategies for disclosure control will need to be better described.
- In the draft document for data access, under adjunct studies, we need to remove details of the data use agreement that have to do with having a released file if the decision is made that no file will ever be directly released to the adjunct study investigator.

1.3 General Scientific Community – Level 3

- Releasing a restricted use microdata file for the general scientific community doesn't really protect confidentiality specifically. It just transfers the responsibility for disclosure control from the government to the individual researcher, usually through a Data Use Agreement. The existence of such an arrangement should be specified in the informed consent.
- The NCS needs to consider what the mechanism will be for distributing microdata to the general scientific community (e.g., downloads? disks? what encryption levels?). Given the decisions made about distribution, what does it mean to destroy or return data, and how would that be enforced?
- For the general scientific community the NCS will need to decide its goals in balancing protections with trusting researchers to adhere to Data Use Agreements.
- The current plan for data release to the general scientific community does not make everything available (moderate statistical disclosure control [SDC] is specified). NCS needs to think about how to justify this plan.
- The Panel recommended that the NCS should consider not having a moderate SDC file for the general scientific community. NCS could have everything available either through released microdata or through a study-wide secure data facility. The suggestion is that at this level (level 4) licensed microdata should be altered by suppression of sensitive data rather than by coarsening. Access to the suppressed data should be through a study-wide secure data facility.
- The risk of targeted re-identification is higher than the risk of random hacking. Hackers are more likely to want to seek information on someone whom they know is participating in the study. There may, however, be some appeal to hackers in general to break into a high-profile study's data.
- Protections against individuals hacking into microdata stored locally by general scientific community investigators can be implemented by requiring that the data be stored on stand-alone, non-laptop machines (no network connection), or by requiring that the machine be disconnected from the network when the data are used, and that the data be cleaned off the machine after use and before reconnection to the network.
- For data for the general scientific community, signed Data Use Agreements might be sufficient to protect the data. One way to make data accessible to the general scientific community is to use very stringent licensing for the data.
- In their Data Use Agreements NCES requires an explanation of the data use, a signature from the PI, and a signature from a senior official at the institution. Every user also signs an agreement. The agreement requires a stand-alone computer (no network connection), and sharing of any publications with NCES. The site may also be subject to unannounced security inspections by the agency, and penalties may be imposed for violations.
- For their Data Use Agreements NHLBI requires three items, and the agreement must be signed not only by the researcher, but by a business official at the site:
    (1) Application stating purpose for using the data
    (2) A storage agreement for the data
    (3) IRB approval

- Another way to protect data is by making people come to a secure data center to conduct all analyses, eliminating distribution of microdata for local storage. NCHS does this with some of their data.
- Deciding what is meant by credentialing will be very important. NCHS does not seek credentials in the form of particular personal qualifications to use data at the level of the general scientific community. They do require a research use (rather than a commercial or administrative use). One example given by a Panel member was of a group that sought data from NCHS about vaccines and was declined access. The group claimed they were being excluded for political reasons, when it was actually for IRB-related reasons. NCHS could combat the accusations because they don't discriminate based on specific scientific, professional, or educational credentials.
- With regard to credentialing, NCES and NCHS don't look at who you are, but may require other credentials like IRB approval of the project.
- The Panel emphasized that we must have full access to the data for anyone who is qualified.
- When NCES evaluates credentials for licensing they do not allow students to be the PI because students are too transient. Faculty advisors must be the PI instead.
- Any researcher should be able to gain access to the full de-identified dataset, even if the NCS has steps of greater security to get greater access. People could start their access with more widely available data to do exploratory analyses and determine the utility of the data for their project. They could then step up the security ladder for access to more complete data, either through a license or through a study-wide secure data facility.
- The NCS might get requests (before people request access to data) about sample characteristics to help people determine whether they have enough representation of their subgroup of interest to conduct analyses. To avoid needing to fill such requests constantly, the NCS might want to consider making basic information publicly available in the form of tables with rounded cells and totals (rounded to protect confidentiality).
- NHLBI puts some of the publications resulting from restricted use (general scientific community) data through a review for disclosure control/confidentiality before publication. They also have optional feedback that they provide on content.

1.4 Public Use Data – Level 4

- The NCS should seek out experts in constructing public use data when we are determining how to construct a useable public use dataset.
- A data analysis system (DAS), which might be a good option for public use files, can be drawn from a restricted use file initially (disclosure control on input) or can limit what can be done or displayed based on data requests (disclosure control on output).
- Because no Data Use Agreement is planned for access to data for the general public, sensitive or identifying data would need to be suppressed.
- It is possible that no useful "public use" data that are available without data use agreements would be feasible. The breadth and sensitive nature of the data might preclude that. If it is not feasible, the NCS might consider doing what NCHS does, which is making data available to the general public through the study-wide secure data facility. In this way the public can access all of the data, but the agency maintains complete control over disclosure risk in the secure setting.

2. Policies

2.1 Data Access and Disclosure Control Policies

- Policies regarding data access and confidentiality must be clear and public in advance of data release. Otherwise external scrutiny will happen once data are released, and people outside the NCS will want access that is faster and more comprehensive than what is indicated.
- Specific policies that are made public up front, including those stated in the informed consent, will make these policies easier to carry out and to defend. Policies made public too far in advance, however, might open the doors to people who want to seek policy changes before the launch of the study.
- The policies that guide access should have explicit principles about dynamic changes in access policy that can occur over the life of the study, and how those changes can occur.
- The data access/use policies need to begin with general principles that guide decisions, but are broad enough to allow for dynamic decisions over time. The NCS can use NCHS principles as a guideline; these can be sent to the NCS by the NCHS Panel member.
- The Panel suggested that the NCS might want to set very high standards for protections initially, but leave room within the policy for the Data Access Committee to make later revisions that might loosen policies. Data access should be a living policy that is revised by the committee as technology changes.
- If the NCS starts out with tight data access policies and loosens them later, the informed consent will need to reflect the broad issues rather than specifics to permit this. The changes to access in such cases might reflect additional data that become accessible rather than changes in who has access.
- The NCS will need to develop a policy for dealing with people who want to match the NCS data to extant data.
- NHLBI differentiates policies for release of data to non-profit vs. for-profit companies (it was not specified during the meeting exactly what is done differently).
- The NCS should consider specifying training as a prerequisite for access to different components of the data. Records can then be kept about who has training to use specific data, and exactly which people have access to which data. Training could be material to read, web-based training, or in-person training.
- Data may need to be apportioned by what everyone can have vs. variables that need to be approved for distribution. Consequently, each data element may need to be evaluated for whether it is sensitive or identifiable, and therefore would require extra approval.
- The Framingham Heart Study (NHLBI) restricts data release, with no data available to the general public, because their data are too identifiable. Similar problems might arise for the NCS.

2.2 Specimen Access Policies

- Although specimens will have separate protections due to their nature as a consumable resource, the NCS will need to determine (and to specify) whether data policies for specimens are the same as for non-specimen data.
- NHANES has a system related to access to specimens that the NCS might want to look at.
- On NCI studies and other similarly sensitive datasets specimen data accessed for a particular "study" are fully anonymized (all links back to the original data are broken). This affords high levels of privacy and confidentiality, but the researchers cannot go back and add new data to their file after anonymization, which makes longitudinal analyses impossible. Also, to use the same sample again it needs to be analyzed anew, even if the same analytes are being examined that were analyzed on the anonymized study, because the links to the rest of the data were broken.

- Anonymizing data on the NCS might be problematic because it would require reanalysis of samples that exist in finite quantities and may have many different uses.
- NCS might want to consider requiring that adjunct studies analyze the full study sample of specimens for whatever they are interested in rather than just analyzing a subset so that specimens don't get analyzed for only a subgroup. It might be possible to direct adjunct study investigators who are interested in a particular analyte for a subset of the NCS study sample to partner with other adjunct study investigators who are interested in the rest of the NCS study sample to result in analysis of the full study sample for the analyte in question. Otherwise specimens may be used up for a select few, and which analytes have been analyzed will be erratic across the study sample.

3. De-identification and Disclosure Control Procedures

- NCS should consider following OMB's guidance to determine the line between de-identification of data and disclosure control. OMB also defines direct identifiers and indirect identifiers separately.
- Exact dates, such as birth dates, should be included in the list of items for de-identification.
- Perturbations such as swapping cannot be done securely at a single level without also doing it at every level. If it's only done for some releases of the data individuals might be re-identifiable through comparisons of the data across releases, or through comparisons between released data and published tables.
- Some information might be masked through coarsening, but such changes should be done in ways that do not affect the validity of the analyses.
- Information disclosed during research activities might have secondary implications for privacy and confidentiality. For example, it would be important to require that presenters avoid identifying segments in articles or conference presentations. Knowledge of segments narrows down the geography too much, and can lead to re-identification of participants. Likewise, some identification of attributes of particular subgroups might narrow down identities of members of those groups.
- Even the existing "blurry maps" of the segments might be too identifying, and should not be distributed or used in public presentations.
- Although the NCS will remove center, site, etc., contextual information and particularly demographic information might still permit re-identification.

4. Legal and Ethical Issues

- Informed consent can be used to protect data and defend decisions, but to do that it must address all levels of data users.
- NCS consent forms are not specific enough. Who will have access is the important element in the consent. The consent needs to specify who can and who cannot get the data (who are "study investigators"). The consent document can be a protective document that NCS can refer to later to restrict access – essentially that NCS promised participants we will only distribute X data to Y people.
- The NCS informed consent might need to make reference to the Privacy Act. The NCS recourse for violations of policies for a license at the general scientific community user level might be able to be tied back to the Privacy Act, although the NCS should see if any other laws protect the data and include enforceable fines.
- Certificates of confidentiality may protect against some things, like individual information sought for divorce or custody cases. It may not, however, protect against Immigration Services or Immigration and Customs Enforcement, or against the U.S. Departments of Justice or Homeland Security. It hasn't been fully tested in court. The NCS should have the DHHS specific language about the certificates in

the informed consent documents, but might want to check with DHHS lawyers about exactly what does get protected.
- The NCS will need to consider intellectual property rights explicitly, specifying how data can be used. This is particularly problematic with drug companies and others who seek to use the data in the creation of patented materials.
- OHRP's standards say that it doesn't have to be impossible to re-identify participants from data; it does, however, have to be very difficult.

5. Data Access Committee

- The Panel recommends that the NCS have a Data Access Committee. That committee might need people external to the NCS to serve on it who can provide an outside perspective on access issues.
- If the agency is ultimately the steward of the data, the Data Access Committee should fall under the auspices of the NCS Program Office rather than under the Steering Committee.
- The Data Access Committee will need to be a standing committee, as issues related to access will be both constant and dynamic.
- The Data Access Committee will be needed to make evolving decisions both about who the data users are (who falls under which level of data access as presented in Table 1), and what exactly researchers at each level have access to.

6. Disclosure Review Board

- The Panel also recommends that the NCS have a Disclosure Review Board, which should not be under the Publications Subcommittee; it needs to maintain independence.
- The Disclosure Review Board should do or oversee the actual work of disclosure control – developing strategies to carry out disclosure control and implementing those strategies.
- The Disclosure Review Board should not set policy. For disclosure issues policy will have to be separated from practice, since policy will be broad and practice may vary by level of user. The policies related to disclosure control should be set at higher levels.
- The Disclosure Review Board may need to review EVERY publication for disclosure risk.

7. Community Relations/Special Sensitivities

- American Indian populations might have different restrictions on specimens and on data distribution than the rest of the NCS study participants, if previous work with tribes by NHLBI is indicative of a general trend. Some tribes won't allow distribution of data or specimens beyond the study investigators. They are concerned about genetic analyses, and about negative attributions to the tribes. Although Peter has been to Apache County to begin outreach and discussions, the reception might be different for each tribal community and should be explored.
- Individual communities might want (or expect) data that pertain to individuals who have certain conditions or live near certain geographic landmarks in their areas (e.g., people who live near power lines in a community who have a particular disorder). Such releases would be highly identifiable because they would often cover only one PSU. The NCS will have to determine how to deal with such requests.
- The NCS should also think about what happens to personal information in the long term, after the original children in the study turn 21. What happens to the contact information for the participants? How will the study deal with requests to re-contact the participants?

8. Miscellaneous

- The Panel reminded the group that release will be expensive, and almost always more expensive than anticipated. Access, release, documentation, and oversight are all very expensive.
- Demand for the data from the general scientific community and the general public will vary with how accessible the data are. Charging for use reduces demand. Pre-conference workshops on the data will increase demand.
- Documentation of the datasets will need to be thorough to permit accurate use of the data.